

# Automatic SNOMED Coding

G. William Moore, M.D., Ph.D., and Jules J. Berman, Ph.D., M.D.

Departments of Pathology, Baltimore VA Medical Center, University of Maryland School of Medicine,  
and The Johns Hopkins Medical Institutions, Baltimore, Maryland

## ABSTRACT

*Medical coding has become an important new industry that has originated from the field of medical informatics. Automatic coding of specimens has emerged as a way of relieving hospitals from the cost of paying professional coders and for achieving uniform coding for all specimens. Unfortunately, automatic coding, like manual coding, has numerous pitfalls. Further, the coding algorithms employed by manufacturers of automatic coders are typically proprietary. We have developed a method for automatic coding of pathology reports. Using this public domain autocoder, we have previously demonstrated that automatic SNOMED coding was superior to manual coding in several measurable categories, including the overall number of codes generated and the number of distinct code entities provided. In this report, we describe an algorithm that executes this strategy in the M-Technology environment.*

## INTRODUCTION

Medical coding has become an industry in its own right. Some hospitals employ professional coders trained to list diagnoses in a manner that supports linkage to reimbursable diagnosis-related groups (DRGs). Inaccurate diagnostic coding may cause a report to be uncountable, irretrievable, or unreimbursable. In the future, coded databases, stripped of patient identifiers and collected from many contributing health care services, may assist epidemiologists in tracking the spread of diseases, identifying areas of special risk, and providing reliable quantitative information for developing national health care policies. All these activities require accurately coded databases (i.e., databases that contain all the codes for all the specimens collected by a pathology department). However, the ability of any pathology department to obtain an accurately coded database is far from trivial.

In one of the few studies addressing the difficulties in coding, Hall and Lemoine [1] found errors in more than 10% of cases. They divided manual coding errors into five types: (1) Factually correct but unhelpful codes (e.g., coding all benign lesions as 'negative for tumor'); (2) Inconsistent codes (coding 'dysplasia' on Monday and 'atypia' on Tuesday); (3)

Idiosyncratic codes (using a mnemonic for a lesion, often inscrutable to other people); (4) Entry errors (e.g., entering 'lipoma' when one intends to enter 'lymphoma'); (5) Incomplete coding due to impatience or laziness.

In our experience, the way that coding is performed varies considerably depending upon the intended use of the codes. For instance, some pathologists attempt to choose the single best code for a given specimen. Other pathologists code a single case under multiple related morphologic or topographic terms to insure the success of some future search. For example, a single vocal cord lesion may be given all the following morphologies and their corresponding codes: cytologic atypia, precancer, dysplasia, carcinoma in situ, squamous carcinoma. The topography code may be listed as larynx, neck, vocal cord, and even respiratory tract. In addition, when the diagnosis of a specimen is equivocal, the pathologist may code for all the possible diagnoses in the histopathologic differential, even when those diagnoses may be mutually exclusive (i.e., reactive atypia and invasive carcinoma). An epidemiologist trying to determine the respective incidences of vocal cord dysplasia and vocal cord carcinoma may be perplexed by the many code listings for a single biopsy specimen. Vendors and pathologists are left on their own to choose a coding strategy, from the two extremes of coding: 1) a single best fit diagnosis or 2) coding all the related terms for a given lesion. This question will have greater relevance when administrators and epidemiologists attempt to collect and use coded databases.

Considering the enormous resources devoted to manual coding, as well as the inescapable introduction of human error into the collected data, the incentives for automatic diagnostic coding are obvious. A variety of software systems that perform automatic coding ('autocoders') are commercially available. Unfortunately, the algorithms, source code and even the basic coding strategies of proprietary systems lie outside the public domain, and have not been scrutinized in the informatics literature. The Systematized Nomenclature of Medicine (SNOMED) is a widely-used coding system in pathology departments [2,3]. In a recent publication, we have compared the performance of a non-commercial (public domain) SNOMED autocoder

against the performance of manual coders [4]. We found that fully automatic SNOMED coding is a practical alternative to manual SNOMED coding, and that automatic SNOMED coding was superior to manual coding in several measurable categories, including the overall number of codes generated and the number of distinct code entities provided. We describe here the techniques we have used in automatic coding and the problems that may arise when coding surgical pathology reports.

## MATERIAL AND METHODS

**Manual Coding.** Manual coding was performed by three board-certified anatomic pathologists at the Baltimore VA Medical Center. Nearly all cases were assigned one topography and one morphology code. The other axes of SNOMED (etiology, function, procedure, disease, occupation) were usually ignored.

**Automatic Coding.** Reports were obtained as a raw global ASCII file downloaded from the mainframe computer at the Baltimore VA Medical Center. The entire contents of each report, including patient demographics, date and time of accessioning and signout, specimen source, gross description, final microscopic diagnosis, pathologist's identification, and manually-entered SNOMED codes, were passed into an ASCII file, a total of 21,168,261 bytes. The full text of the 'specimen source' and 'final microscopic diagnosis' for each case served as source text for the SNOMED autocoder. All numerals, punctuation marks, and barrier words (see below) were removed from the source-text, as well as all letter-strings shorter than 3 letters, except for: 'no', 'os' (= 'bone' or 'left eye'), 'od' (= 'right eye'), 'eg' (= 'esophago-gastric'), and 'ge' (= 'gastro-esophageal'), to produced a REDUCED REPORT, as shown in the following examples. The manual/autocoder discrepancies in these examples are typical of our experience.

### EXAMPLE 1:

#### ORIGINAL REPORT:

SPECIMEN: 1. TOE.

DIAGNOSIS: 1. BONE WITH HYPERTROPHY AND GOUTY TOPHUS.

#### REDUCED REPORT:

toe bone hypertrophy gouty tophus

#### MANUAL CODES:

TY9800 TOE

M71000 HYPERTROPHY

#### AUTOMATIC CODES:

TY9800 TOE

T1X500 BONE

M71000 HYPERTROPHY

### M55070 TOPHUS

#### EXAMPLE 2:

##### ORIGINAL REPORT:

SPECIMEN: 1. PUNCH BIOPSY RIGHT GROIN.

DIAGNOSIS: 1. CHANGES CONSISTENT WITH MILD, SUBACUTE DERMATITIS.

COMMENT. THERE IS HYPERKERATOSIS, FOCAL PARAKERATOSIS WITH CRUST FORMATION, ACANTHOSIS, MILD SPONGIOSIS, AND MILD UPPER DERMAL PERIVASCULAR CHRONIC INFLAMMATION. PAS STAIN IS NEGATIVE FOR FUNGI. PSORIASIS AND KAPOS'S SARCOMA ARE NOT LIKELY, AND THERE IS INSUFFICIENT SPONGIOSIS FOR A DIAGNOSIS OF SEBORRHEIC DERMATITIS.

##### REDUCED REPORT:

right groin subacute dermatitis hyperkeratosis parakeratosis crust formation acanthosis spongiosis dermal perivascular chronic inflammation negative fungi psoriasis kaposi sarcoma not spongiosis seborrheic dermatitis.

#### MANUAL CODES:

T01000 SKIN

M43000 CHRON. INFLAM.

#### AUTOMATIC CODES:

T01000 SKIN

TY9800 INGUIN. REGION

M36380 CRUST

M72600 HYPERKERAT.

M42000 SUBAC. INFLAM.

M43000 CHRON. INFLAM.

M30490 FUNGUS BALL

M74030 PARAKERAT.

M36500 EDEMA

M48840 PSORIASIS

M91403 KAPOS'S SARC.

M48820 SEBORR. DERM.

### EXAMPLE 3:

#### ORIGINAL REPORT:

SPECIMEN: 1. LENS OD.

DIAGNOSIS: 1. CATARACT (GROSS DESCRIPTION ONLY).

#### REDUCED REPORT:

lens od cataract

#### MANUAL CODES:

TXX700 LENS, NOS

M51100 CATARACT

#### AUTOMATIC CODES:

TXX756 LENS, RIGHT

M51100 CATARACT

TXX700 LENS, NOS

### EXAMPLE 4:

#### ORIGINAL REPORT:

SPECIMEN: 1. RIGHT PROSTATE BIOPSY.

2. LEFT PROSTATE BIOPSY.  
 DIAGNOSIS: 1, 2. BENIGN STROMAL AND  
 GLANDULAR HYPERPLASIA.

REDUCED REPORT:  
 prostate prostate benign stromal glandular hyperplasia

MANUAL CODES:	AUTOMATIC CODES:
T77100 PROSTATE	T77100 PROSTATE
	M09450 NO MALIGN.
M72000 HYPERPLASIA	M72400 STR. GL.HYP.

EXAMPLE 5:

ORIGINAL REPORT:  
 SPECIMEN: 1. PUNCH BIOPSY FOREHEAD.  
 DIAGNOSIS: 1. ACTINIC KERATOSIS.

REDUCED REPORT:  
 forehead actinic keratosis

MANUAL CODES:	AUTOMATIC CODES:
T01000 SKIN	T01000 SKIN
	T10110 FOREHEAD
M72850 ACT.KERAT.	M72850 ACT. KERAT.

In our SNOMED autocoder, a word-sequence of arbitrary length in each pathology report is pointed to a one-or-more SNOMED codes in the dictionary. Each SNOMED code is repeatedly enriched with additional synonyms, using the 'barrier word method', as described below. In this manner, nearly every significant term in the three years of reports issued by our department could be captured and pointed to an appropriate SNOMED code in a timely fashion. This approach requires one person to function as a 'dictionary policeman' within the department, but is repaid by a very low level of false negatives.

Automatic coding of free-text diagnoses into SNOMED codes was performed on TRANSOFT, a table-driven public-domain computer translation shell, written in M or HyperPAD [5,6]. TRANSOFT is designed for translation between any two languages using the Roman alphabet. The M source code is available through Internet [7]. TRANSOFT is embedded in the File Manager (FileMan), the core database management and program development environment of the Decentralized Hospital Computer Program of the U. S. Department of Veterans Affairs [8]. The user supplies the dictionary and a grammar in the augmented transition network style, which is

common to many computer translators [9]. Input is through the FileMan user interface or through an ASCII word processor. The user controls the behavior of the translator through externalized language-specific information and generic program code. TRANSOFT prototype translators have been constructed between English and several languages, simply by changing the FileMan databases [5].

Barrier Word Method. The 'barrier word method' is a computer method for extracting multiple-word terms from a free-text document. All punctuation-marks, numerals, articles, prepositions, and common adjectives and verbs are called 'barrier words' (alternatively, 'stop words'). In addition, each large source document in a particular subject area will have its own, idiosyncratic set of barrier words, which become apparent after repetitive application of the barrier word method to that document. For example, the following report, the barrier-words are shown in lower-case and the remaining, main-words are shown in upper-case:

specimen: 1 . biopsy APPENDIX .  
 2 . biopsy CECUM .  
 3 . biopsy HEPATIC FLEXURE .  
 diagnosis: 1 . COLONIC MUCOSA with rare CRYPT  
 ABSCESS and CRYPTITIS .  
 2 . COLONIC MUCOSA with focal  
 ULCERATION , FIBRINOPURULENT MATERIAL ,  
 and GRANULATION TISSUE .  
 3 . COLONIC MUCOSA with mild EDEMA  
 and rare NEUTROPHILS .  
 comment: findings are consistent with mild  
 INFLAMMATORY BOWEL DISEASE .

barrier words:	MAIN WORDS:
specimen	APPENDIX
biopsy	CECUM
diagnosis	HEPATIC
with	FLEXURE
rare	COLONIC
and	MUCOSA
with	CRYPT
focal	ABSCESS
and	CRYPTITIS
with.....	COLONIC.....

It is apparent from this short excerpt that many multiple-word-sequences of main words that appear between two consecutive barrier words constitute a technical term that might possibly be pointed to a SNOMED code, as follows:

#### MULTIPLE-WORD TERMS:

HEPATIC FLEXURE  
COLONIC MUCOSA  
CRYPT ABSCESS  
COLONIC MUCOSA  
FIBRINOPURULENT MATERIAL  
GRANULATION TISSUE  
COLONIC MUCOSA  
INFLAMMATORY BOWEL DISEASE

False-negative and False-positive rates. A 'false-negative case' is one to which a correct code for a major diagnosis has not been assigned. A 'false-positive case' is one to which an incorrect code for a major diagnosis has been assigned. The 'false-negative rate' is the proportion of false-negative cases among all cases. The 'false-positive rate' is the proportion of false-positive cases among all cases. In principle, false-negative and false-positive rates may be obtained both for manual coding as well as for the various methods of autocoding. Unfortunately, obtaining these rates requires that each case be examined by a human coding expert, and the correct codes determined for that case. From this set of 'true positive' codes, a computer program can determine whether a particular case has been correctly assigned by manual or various automated methods. Most pathology laboratories cannot devote the human resources necessary to determine the exact set of true-positive codes for their caseloads.

For retrieval problems, the most important information is the false-negative rate for the autocoder. This is the proportion of cases in which the autocoder fails to assign a correct code needed for retrieval. If the autocoder has, say, a 10% false-negative rate, this means that, on average, 10% of cases desired in a particular retrieval request will not be recovered. The false-positive rate, namely the proportion of unwanted cases that will be recovered, can be regarded as a nuisance-factor, which only becomes important if it is very large. For example, when one performs a MEDLINE literature search, one typically detects numerous unwanted citations; but these can easily be bypassed at a glance. The desired citations which are not detected (false-negatives) is the more vexing aspect of a literature search.

For the present investigation, we assumed initially that the manual coding for each case contained no false-negatives for major diagnoses. That is, we assumed that the major sense of the case was always captured manually. We then reviewed every case in which a major diagnosis from manual coding had been missed by

the autocoder. The list of 'major missed diagnoses' was obtained as follows: First, we assembled a list of 'minor diagnoses', such as 'M09450 NO EVIDENCE OF MALIGNANCY', 'M00100 NORMAL TISSUE MORPHOLOGY, NOS', as well as non-specific inflammation, such as 'M41000 INFLAMMATION, ACUTE, NOS', 'M43000 INFLAMMATION, CHRONIC, NOS', etc. A minor diagnosis in the manual coding was not required to find a match in the autocoder diagnoses. Second, a list of near-synonyms was assembled, such as 'M81400 ADENOMA' near-synonym for 'M82110 TUBULAR ADENOMA'. A major diagnosis in the manual coding was considered matched if its near-synonym appeared in the autocoder diagnoses. Finally, a match was only required in the first three digits of the SNOMED-code (where the first digit is either 'M' or 'T'). Thus, 'M72000 HYPERPLASIA' was considered a match for 'M72400 HYPERPLASIA, GLANDULAR AND STROMAL'.

## RESULTS AND DISCUSSION

A total of 9,353 cases was examined over the 33-month duration of the study [4]. In the first pass of the autocoder, 463 (5%) discrepant cases were detected, in which a major diagnosis in the manual coding had been missed by the autocoder. A final set of true-positive diagnoses were assigned to the initially discrepant cases, and the cases were passed through the autocoder again. In this second pass, there was a missing, major, true-positive diagnosis in only 44 (0.5%) cases.

The nomenclature for automatic coding is somewhat vague. The term 'computer-assisted coding' has been used to refer to a variety of distinctly different activities. Our impression is that the term 'computer-assisted coding' describes a system where the person entering data is prompted by the computer to enter the name of a topographic site or morphologic entity. The computer then points to a matching entry, if any, in the SNOMED file. If there is a match, then the computer reports the code number assigned to the matching file entry. If there is no match, then the user is prompted to enter another morphologic diagnosis or topography. It is our experience that most pathologists regard this form of coding as 'manual' coding, since the pathologist must manually re-enter the specimen source and final microscopic diagnoses for every specimen. This system is faster than searching for diagnoses in the SNOMED books, but is not as fast as having the computer extract codes from the free text

report. We use the term 'automatic coding' to describe systems in which the computer does all of the work of coding, with no user interaction.

Confusion with these aspects of SNOMED coding is reflected in the complex strategy that we finally settled upon for comparing manual coding to results of the autocoder. First, we assembled a list of 'minor diagnoses', such as 'M09450 NO EVIDENCE OF MALIGNANCY', which were not required to find a match among the autocoder diagnoses. Second, a list of near-synonyms was assembled, such as 'M81400 ADENOMA' near-synonym for 'M82110 TUBULAR ADENOMA', in which the manual coding was considered matched if its near-synonym appeared among the autocoder diagnoses. Third, a match was only required in the first three digits of the SNOMED-code, so that, say, 'M72000 HYPERPLASIA' was considered a match for 'M72400 HYPERPLASIA, GLANDULAR AND STROMAL'. Finally, we found it necessary to have a 'dictionary policeman', who reviewed all new encounters with previously unused phrases occurring in our natural language text file, and pointed these phrases to appropriate SNOMED codes. Without these conditions, the performance of the autocoder would have been appreciably worse. This experience suggests that many departments which employ autocoders will have significant deterioration in performance, unless the autocoders are continually updated.

Remarkably, this automatic SNOMED coding strategy resulted in only 0.5% missed major SNOMED codes by the autocoder as compared to the spell-corrected manual codes. The missed major codes were the result of complex syntax in the source text stream, which would require a sophisticated parsing algorithm [5]. This result suggests that perfect orthography in the source text and vigilant dictionary maintenance are sufficient to achieve highly accurate coding. Complex parsing algorithms, available in computer translators such as TRANSOFT, could not be expected to increase coding accuracy to an appreciable extent.

Currently, coding in pathology departments is done primarily so that reports of a certain lesion or location can be recovered by the pathologist. In the near future, coding activities may relate more closely to broader questions of regional, national, and international importance. Once uses of coded reports become prioritized, and an optimal coding dictionary can be chosen. Additionally, coding algorithms can be designed to minimize errors based on the intended uses of the codes.

## REFERENCES

1. Hall PA, Lemoine NR. Comparison of manual data coding errors in two hospitals. *J Clin Pathol* 1986;39:622-626.
2. College of American Pathologists. Systematized nomenclature of medicine (SNOMED), College of American Pathologists, Skokie, 1976
3. Cote RA, Robboy S. Progress in Medical Information Management: systematized nomenclature of medicine (SNOMED). *JAMA* 1980; 243:756-762.
4. Moore GW, Berman JJ: Performance analysis of manual and automated systematized nomenclature of medicine (SNOMED) coding. *Am J Clin Pathol* 1994; 101:253-256.
5. Moore GW, Wakai I, Satomura Y, Giere W. TRANSOFT: Medical translation expert system. *Artif Intell Med* 1989;1:149-157.
6. Moore GW, Berman JJ. Object-oriented English-to-SNOMED translator using TRANSOFT+HyperPAD. 15th Annual Symposium on Computer Applications in Medical care. 1991; 15:973-975, Washington, DC.
7. TRANSOFT source code (in the M language, ISO standard 11456) may be obtained through anonymous ftp at nctucca.edu.tw, pathname /misc/medicine/transoft, filename trs.zip. Fax inquiries to: 1-410-433-6324.
8. Davis RG: FileMan: A User Manual. National Association of VA Physicians, Bethesda, 1987.
9. Woods W: Transition network grammars for natural language analysis. *Commun Assn Comp Mach* 1970;13:591-606.
10. Moore GW, Miller RE and Hutchins GM. Indexing by MeSH titles of natural language pathology phrases identified on first encounter using the barrier word method. In: Computerized Natural Medical Language Processing for Knowledge Engineering, JR Scherrer, RA Cote and SH Mandil (eds.), Elsevier Science Publishers, North-Holland, pp 29-39, 1989